DOCUMENT RESUME

ED 159 183 TM 007 367

AUTHOR Roid, Gale; Haladyna, Tom
TITLE A Comparison of Several Multiple-Choice, Linguistic-Based Item Writing Algorithms.
PUB DATE Mar 78
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Algorithms; *Cloze Procedure; Comparative Analysis; Complexity Level; *Criterion Referenced Tests; Higher Education; *Item Analysis; *Multiple Choice Tests; Reading Comprehension; Secondary Education; Standard Error of Measurement; Structural Analysis; *Test Construction; Test Items; Test Validity

ABSTRACT

The technology of transforming sentences from prose instruction into test questions was examined by comparing two methods of selecting sentences (keyword vs. rare singleton), two types of question words (nouns vs. adjectives), and two foil construction methods (writer's choice vs. algorithmic). Four item writers created items using each combination of the methods listed above, for a 1,000 word passage. A sample of 273 college and high school subjects were randomly assigned to pretests and posttests containing a balanced mixture of items of each type written by each item writer. Dependent measures were (1) the pattern of pretest and Posttest item difficulties; (2) the variability of item difficulties across item writers; (3) the standard error of measurement of scales composed of items of each type; and (4) the concurrent validity of these scales. No significant differences between item writers were found in the analysis of item difficulties. Keyword nouns, which are relatively rare words in American textbooks but occur very frequently in a prose passage, were found to be unacceptable as question words. Algorithmic methods of foil writing were found to be feasible. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED159183

A Comparison of Several Multiple-Choice, Linguistic-Based Item Writing Algorithms[1]

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gale Roid

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Gale Roid and Tom Haladyna

Teaching Research Division
Oregon State System of Higher Education
Monmouth, Oregon

Paper presented as part of the symposium "Item Writing for Domain-Referenced Tests: Recent Developments" at the meetings of the American Educational Research Association, Toronto, March, 1978.

TM007 367

ABSTRACT

An emerging technology of item writing has been developing in recent years. This study examines the technology of transforming sentences from prose instruction into test questions. The study compared two methods of selecting sentences (keyword vs. rare singleton), two types of question words (nouns vs. adjectives), and two foil construction methods (writer's choice vs. algorithmic).

Four item writers created items using each combination of the methods listed above, for a 1000 word passage. A sample of 273 college and high school subjects were randomly assigned to pretests and posttests containing a balanced mixture of items of each type written by each item writer. Dependent measures were (a) the pattern of pretest and posttest item difficulties, (b) the variability of item difficulties across item writers, (c) the standard error of measurement of scales composed of items of each type, and (d) the concurrent validity of these scales.

No significant differences between item writers were found in the analysis of item difficulties, indicating that the methods are apparently not susceptible to item-writer bias. Keyword nouns, which are relatively rare words in American textbooks but occur very frequently in a prose passage, were found to be unacceptable as question words. Algorithmic methods of foil writing were found to be feasible.

A Comparison of Several Multiple-Choice, Linguistic-Based Item Writing Algorithms[1]

The most widely employed methods of writing test items, for both criterion and norm-referenced tests, rely on the intuitive skills of the item writer who writes items to match objectives or panels of experts who judge the merits of questions. Many theorists, among them Bormuth (1970), Cronbach (1970) and Millman (1974), have convincingly argued that there is a need for a more scientific basis for writing achievement test items. These theorists are on firm empirical grounds when they conclude that two writers using the same learning objectives will not necessarily write similar items. In a study by Roid and Haladyna (1978), one item writer consistently produced more difficult items despite the fact that each used the same set of objectives. This type of bias could result in test forms of different difficulty which would have undesirable consequences on students who take the more difficult tests.

A technology of item writing, involving the use of algorithms for generating test items, could eliminate the bias attributed to item writers by providing methods that are

1. based on a logical and precisely-defined relationship between the instruction given and the test items written,
2. defined by a set of operations open to public inspection rather than defined by the private and intuitive standards of the item writer, and
3. easily replicated by other test developers and not just the original item writer.

Item writing methods with these attributes have the promise of leading to more precise instruments and increasing the accuracy of measurement. A

technology of item writing would contribute toward improving instructional research, educational evaluation, and the use of test data in the formation of public policy (Bormuth, 1970).

There is also a practical reason for interest in algorithmic methods of writing test questions. A large number of items are needed, particularly for instructional systems involving the mastery learning model (Bloom, 1968), in order to provide multiple test forms for use by students who may be retested several times. Algorithmic methods for creating items that are instructionally relevant are therefore very desirable because of the potential for producing large numbers of useful test items.

Algorithm-based Theory of Item Writing

Bormuth (1970) proposed operationally-defined item writing rules for achievement tests used to assess learning from prose material. These rules are a series of directions which tell an item writer how to rearrange segments of the prose instruction to obtain the items of that type. Bormuth described two general types of transformations that are possible on prose material, (a) items derived from sentences and (b) items derived from the relationships between sentences (1970, pp. 39-55). An example of sentence-derived items that test recall of prose material are those produced by the "wh-transformation." Wh-transformation items would be written using a detailed rule summarized as follows: "Inspect all sentences in the instruction, replacing a word such as who, what, or where for the appropriate part, e.g., subject, in each sentence." For instance, "The boy rode the horse" could be transformed to: "Who rode the horse?" Wh-items are particularly useful because they can be written to question

each separate part of a sentence, and can be made into either multiple-choice or fill-in format.

Finn (1975) has extended the work of Bormuth by developing a rather detailed algorithm that originally contained 82 steps. For purposes of the present research, the original algorithm has been streamlined (Finn, Note 1; Roid, Haladyna and Finn, Note 2).

To develop questions that measure important aspects of a prose passage requires, as a first step, an objective analysis of the text. The approach proposed by Finn (Note 1), involves a computer analysis of the words in a prose passage. All the words in the passage are key-punched for input to a computer program that does two major tasks: (1) counts the number of times that each work appears in the passage and (2) identifies the standard frequency index (SFI) of each word. The SFI of each word is a numerical estimate of how often the word occurs in a large corpus of American English (Carroll, Davies & Richman, 1971). SFI ranges from 88.6 for the word "the" (meaning that the average American student is likely to encounter the word "the" once in every ten words of his school book reading) to 02.5 for the word "incarnation" (meaning that the average student is likely to encounter the word "incarnation" less often than once in every billion words of his school book reading).

The objective of this analysis is to identify "high information" words--words that are relatively rare in American English and occur only a single time in the passage. The sentences in which these high information words occur can then become candidates for transformation into questions which tap important information in the passage. High information words are those that are difficult for subjects to guess if they are

deleted from a prose passage such as is done in a Cloze test (Culhane, 1970). The easiness with which a word is guessed or filled in by a subject on a Cloze test is a measure of the amount of information it provides. The information in a missing or garbled message is a function of the amount of doubt the receiver has about its identity and is related to the probability of occurrence of certain words or letters. A missing word which is highly predictable or a common word in the English language would present less information because subjects would have less doubt about the possibility that it occurs in the missing place.

Finn (Note 1) has done considerable research to show that the Cloze easiness of a word is predicted by the two important indexes derived from the computer analysis of a passage: (1) the standard frequency index (SFI) and (2) text frequency. Words that have a low SFI (infrequent in American school books) are typically high in information. However, there is one circumstance in which the information of these words is reduced in relation to a given passage. If the word is repeated frequently (i.e., it has a high text frequency), the information value of that word is diminished because subjects will supply it more often in a Cloze test following reading of the passage. In other words, repetition of words, even if they are rare in American English, lowers their information value. Therefore, Finn has proposed that candidates for good question words are those which are both rare in American English (have a low SFI) and are singletons in a prose passage (i.e., have a text frequency of 1).

Purpose

The purpose of this study was to examine test items created by variations in algorithms. This research focuses on the multiple-choice

item and its two important parts: the stem and the options. The stem is that section of the question which requests a response from the student and the options include the correct answer and the various wrong-answer alternatives (foils). Specific research questions of the study were:

1. What method for identifying a keyword to be selected from a prose passage is most useful? Specifically, how do "high information" words (rare singletons) compare to rare words with a high text frequency (keywords)?
2. How do noun-based items differ from adjective-based items?
3. How do algorithmic methods of option construction compare to methods in which each item writer selects the options?
4. How do different item writers vary in their choice of options and in their implementation of algorithmic item writing methods?

Method

Design

Four item writers created items in each of the following categories which serve as the independent variables for the study: (a) items based on two stem-generating techniques--rare singletons and keywords, (b) noun-based items and adjective-based items, and (c) items based on two foil construction techniques--algorithmic and writer's choice. Four dependent measures were used. The first two involve item difficulties, while the latter two involve measures of reliability and concurrent validity that pertain to scales which are composed of items representing a particular technique, item writer or combination thereof.

The first of these dependent variables was instructional sensitivity, which is the tendency for items to have low difficulty indexes when administered to non-instructed students and high difficulty indexes when administered to instructed students. Difficulty is the percentage of items correctly answered by a group of students. In instructional settings, instructional sensitivity is one way in which to validate an item as a measure of the general trait being developed for which instruction is planned (Haladyna & Roid, Note 3).

The second dependent variable is the variability of item difficulties for items which are grouped by the independent variables of the study. Bormuth (1970) has hypothesized that if objectivity in item writing is achieved, then the variability of items produced by various item writers should be small rather than what is typically observed when item writers produce items using intuitive methods. It follows that algorithm-generated items, when so grouped, should have less item difficulty variability than items produced in other ways.

A third indicator of effectiveness of a particular item writing approach is the degree of measurement error of a scale based on a particular set of items representing any level of any independent variable. For example, standard errors of measurement of a scale representing rare singleton items were compared to the standard errors of a scale composed of keyword items.

A final measure of effectiveness is the relationship of each scale with a criterion measure, Cloze scores. The Cloze score yields objective measures of comprehension which are not influenced by item writer bias or guessing. A technique or combination of techniques (e.g., keyword,

noun-based, algorithm-generated) will have various degrees of concurrent validity, the Cloze score serving as the criterion by which we evaluate these scales. A second type of concurrent validity is the correlation between scale scores representing techniques or combinations of techniques and the total test score. This is an internal type of concurrent validity, while the Cloze test can be considered external to the test situation.

Item Development

Patrick Finn and three graduate students from the State University of New York at Buffalo analyzed the text and produced the questions for the research study. This group worked as a team in producing several trial questions on the prose passage in order to be sure that close agreement was made as to the classification of words and transformation of sentences. After this initial team work, the writers worked independently in producing questions.

The important steps in the algorithmic development of items were (a) analyzing the text, (b) selecting the question words and sentences, (c) preparing sentences, (d) transforming sentences into item stems, and (e) generating options. For further details of these methods, see Roid, Haladyna and Finn (Note 2).

Analyzing text. A prose passage, at the approximately high school level of reading, was selected on insect development. All words in the passage were subjected to a computer analysis that identified the text frequency and the standard frequency index of each word in the passage. Words with a SFI (Carroll, Davies & Richman, 1971) of 60 or less were selected for possible question words in the categories of adjectives and

nouns. Among all of the adjectives or nouns with a SFI of 60 or less, two major types of words were selected: (1) words having a text frequency of 1 and (2) words having the highest text frequency. Words with a text frequency of 1, which also have a SFI of 60 or less, are considered to be high information words. In the case of the words which had a high text frequency, it was necessary to select a particular occurrence of the word in the text. For purposes of the present research an algorithm was constructed based on the choice of the first occurrence of a high text frequency word. These words are referred to as "keywords," and high information words as "rare singleton" words.

Selection of question words and sentences. Not all parts of speech are equally good candidates for question words even though they may be high information words. After attempting to produce questions from verbs, Roid, Haladyna and Finn (Note 2) concluded that the most promising parts of speech are adjectives and nouns.

For these reasons, five sentences from the passage were selected for transformation into items in each of the following categories: (1) keyword adjectives, which are adjectives with a SFI of 60 or less but which had a high text frequency, (2) keyword nouns, which are nouns with a SFI of 60 or less but had a high text frequency, (3) rare singleton (text frequency = 1) adjectives, and (4) rare singleton nouns. This represents a total of 20 sentences from the prose passage. Each sentence was transformed by each of four item writers who used two foil methods, one based on an algorithm and the other simply being the item writer's choice (a non-algorithmic method). A total of 160 items were the objects of this analysis.

Preparation and analysis of sentences. Once an important word had been identified for a possible question, the sentence it occurred in was examined and prepared for transformation. The important parts of this transformation were the identification and elimination of non-functional imbedded sentences, the replacement of indefinite pronouns with the phrase or word being referred to in a previous sentence, and the actual rearrangement of words in the sentence.

Transformation of the sentence into a question. The next step in processing the sentence was to eliminate the question word and to transform the sentence into a question. The question word, and the entire noun phrase in which it occurred, was removed, and was replaced with a wh-word. Where several wordings were possible, an attempt was made to stay as close to the wording of the original sentence as possible.

The algorithm does not produce 100% agreement among item writers in all cases because of such things as the replacement of phrases from previous sentences and Finn (1975, pp. 357-363) discusses some of the discrepancies among item writers.

Generating foils. As is common knowledge among test developers, the writing of good foils for multiple-choice questions is challenging and important work. The first step in an algorithmic generation of foils is to classify the question words which are the correct answers to the question so that possible foils can be obtained from a list of words in the same classification. The most logical source of foil words would seem to be from the prose passage itself, and the present study involved the development of methods for doing that.

Two variations of the algorithm were developed: one for nouns and one for adjectives. In the case of nouns, all nouns with a standard frequency index of 60 or less were semantically classified using the method of Fredericksen (1975). For example, some nouns were classified as concrete animate nouns. For a given question word, a random sample of other nouns from the passage that were in the same semantic classification were drawn to create foils.

In the case of adjectives, research on semantic differential technique was used as a basis for constructing a classification of adjectives from the passage (Nunnally, 1967, pp. 536-538). In semantic differential research there are typically three or four factors that are descriptive of words: (1) evaluation, such as "good," "bad,", etc., (2) potency, such as "strong," "weak,", etc., (3) activity, such as "fast," "slow," and (4) familiarity, such as "simple," "complex." These four classifications were used to group the adjectives in the passage that had SFI's less than 60. These adjectives were then subjected to an analysis of whether or not they were "familiar," as determined by the Dale-Chall List of 3,000 Familiar Words (Dale & Chall, 1948). The adjective, to be a candidate for a foil, needed to be absent from that list. The rationale for this was that there are some extremely common adjectives that would be too easy as foils.

From the insect development passage there were approximately 50 adjectives which passed this screen and were qualified as possible adjectives for foil construction. A fifth classification of adjectives was used in this study. Particularly in prose material that is technically oriented, a "technical word" category of adjectives can be useful

in grouping those words which are very specific to a subject matter in a particular passage. This was done with reasonably good agreement between judges. In the case of the insect development passage, technical words were those which were of scientific origin describing insect development. To build foils for adjective question words, a random sample of adjectives was drawn from the classification category of similar adjectives.

The algorithmic method of foil construction was compared, in the present research, to a more informal method where each item writer chose their own foils. Each writer inserted a word of their choice into the phrase containing the question word. No other words in the phrase were changed by the writer. For example, for the sentence, "Each egg produces a single immature insect," the question word was "immature," and the phrase deleted and made into foils was "a single _____insect." Each writer inserted a new word into the phrase for each foil, e.g., "a single nymphal insect."

Even when algorithmic foils are used, the identity of the writer may show in the choice of wording that produced the stem of the question. For this reason, the writer and foil methods are true crossed factors in this experimental design. As described above, separate methods of classifying foils were used for nouns and adjectives.

Test Development

The total of 160 items were equally distributed among the independent variables of the study: (a) rare singleton words and keywords item stems, (b) noun-based and adjective-based item stems, (c) algorithm-based and writer's-choice foils, and (d) four item writers. Eight test forms

were created by assigning groups of items representing equally all independent variables to each form. Thus, each test form was balanced with respect to all independent variables as well as the 20 sentences from the passage upon which items were based.

A Cloze test was developed by taking a random sample of paragraphs from the text, and deleting every fifth word of running text. Then, the subject was to fill in the blank spaces left by deleted words. The total length of the Cloze passage was 262 words with 50 deleted words.

Testing Procedures

A sample of 24 college students and 249 high school students were randomly assigned to one of eight 20-item test forms for a pretest. They then studied the prose passage and were given a posttest form and the Cloze test. Approximately thirty subjects were randomly assigned to each test form on the pretest and the posttest interaction.

Data Analysis

1. Instructional sensitivity. To examine the first dependent variable in the study, a repeated-measures analysis of variance of the item difficulties was used to compare items of each type for each test form. A 4 x 2 x 2 x 2 x 2 factorial design was used to examine differences in item difficulties between (1) the four item writers, (2) adjective-based and noun-based items, (3) keywords and rare singletons items, (4) the two foil types (writer's choice and algorithmic), and (5) the two test occasions (pretest and posttest). The analysis of variance technique is useful for determining the "instructional sensitivity" of items through any interaction with the repeated measures, pretest and posttest.

2. Variability of item difficulties. To examine the possibility that item writers using algorithms produce items with less variability with respect to difficulty, a repeated-measures analysis of variance in which the dependent variables were the natural logarithms of the variances was used (Scheffé, 1959, p. 83). The procedure parallels the conventional analysis of variance with the major exception being that the variances of item difficulties for the cells of the design are subjected to analysis to determine if an item generating technique or combination of techniques is producing greater or smaller variability of item difficulty indexes.

3. Standard errors of measurement. Standard errors were computed in the conventional manner, and provided an index of the precision of measurement for scales representing each type of question. Internal consistency reliability estimates were also examined, but fluctuated widely between test forms and provided an unclear view of the differences among question types. Standard errors, on the other hand, didn't fluctuate in this manner, and provided a clearer view of differences between methods. Also, the standard error is very useful in instructional applications of tests, particularly mastery testing, where decisions of a pass/fail nature are made based on test scores.

4. Concurrent validity. Finally, the concurrent validity of each scale was assessed by correlating scores from each scale with Cloze scores and with total scores. Better measures of the underlying trait being modified through instruction should be manifest through higher correlations with these Cloze scores and total test scores and not be influenced by the particular item writer.

Results and Discussion

1. Instructional Sensitivity. Six statistically significant effects were found in the analysis of variance on item difficulties: (a) a main effect for the repeated-measure factor (R) test occasions (pretest vs. posttest), $F(1,128) = 472.03$, $p < .001$; (b) a main effect for part of speech (P) of the question word (noun vs. adjective), $F(1,128) = 12.99$, $p < .001$; (c) an interaction (PS) between part of speech and stem type (keyword vs. rare singleton), $F(1,128) = 14.21$, $p < .001$; (d) an interaction (RP) between the repeated measure and part of speech, $F(1,128) = 4.76$, $p = .03$; (e) an interaction (RFS) between the repeated measure, foil type, and stem type, $F(1,128) = 9.25$, $p = .003$; (f) an interaction (RPS) between the repeated measure, part of speech, and stem type, $F(1,128) = 20.42$, $p < .001$. Cell and marginal means and standard deviations for these effects are presented in Table 1.

Insert Table 1 about here

The strongest effect was the main effect for test occasions indicating that, across all types of items the percentage of subjects getting pretest items correct was lower than the percentage of subjects getting posttest items correct. In other words, most items showed that the instruction was effective and the tests sensitive to the learning that occurred. Pretest item difficulties averaged 47.6% and posttest item difficulties averaged 74.4% across all items as shown in Table 1. This indicates that the subjects did learn from the passage. With four-option multiple-choice items such as those used in this study and no previous

knowledge of the subject matter, excellent items should show pretest difficulties nearer to the level of random guessing (25%).

Two important findings were the main effect for word type (P) and the interaction of word type and the repeated measure (RP) as shown in Table 1. An inspection of average item difficulties for items constructed from each word type, as presented in Table 1, reveals that items based on question words that were nouns were significantly easier (65.6%) overall as compared to adjective items (56.3%). Table 1 reveals an ordinal interaction between the repeated measure factor (R) and part of speech (P) which indicates that noun-based items had greater instructional sensitivity than adjective-based items. In other words, the difference between pretest and posttest difficulties was significantly greater for nouns than for adjectives (+29.5% vs. +24.1%).

An examination of the PS and RPS interactions, as shown in Table 1, reveals further the source of the difference between nouns and adjectives in this study. It is the keyword nouns that result in the easiest questions (average difficulty 72.4%), as indicated by a Newman-Kuels *a posteriori* analysis of the means in Table 1 (PS interaction). Keyword nouns are nouns from the passage having a SFI of 60 or less but which have a high text frequency. The first occurrence of the keyword in the text was used to create the item. An examination of the sentences in which these keyword nouns appeared provides a possible explanation for the results. These sentences are typically introductory sentences which are very general and which address the main topics of the entire passage. For example, in the passage on insect development, the keyword noun "insects" appears in the very first sentence of the passage and is a very

general statement, "The life of most insects is short but active." Questions derived from this type of sentence seem to be invariably a common knowledge type of item which students can answer without having read the prose passage. Also, keyword nouns were relatively easy for subjects to recall on the posttest (average item difficulty of 83.5%). This may be due to the fact that these keywords were mentioned several times in the passage. Easiness of these items on the posttest is in line with Finn's hypothesis (Note 1) that the information content of these rare words is reduced by their high text frequency.

Finn's hypothesis concerning reduced information content of keywords would seem to be inconsistent with the fact that keyword adjectives produced the most difficult items (53.4% as shown in Table 1). However, an examination of each of the keywords provides a rationale for this inconsistency. The keywords and their text frequency are given in Table 2.

Insert Table 2 about here

Table 2 clearly shows that the keyword adjectives had much lower text frequency than keyword nouns. Thus, Finn's hypothesis applies in that higher text frequency was related to the easiness of items constructed from keywords. With text frequencies of 2 or 3, the keyword adjectives were very close to being rare singletons. Keyword nouns included "insect" and its plural "insects," having a total of 28 occurrences in the text. Also, plurals of "egg" and "adult" appeared 15 and 6 times, respectively, in the passage, to further increase the familiarity of these words to the reader.

The rare singleton nouns (instars, cicadas, silverfish, wasps and appetites) showed a good pattern of pretest and posttest difficulties. They had the highest average instructional sensitivity (40.4% to 77.3%), a difference of 36.9% as shown in Table 1. The rare singleton adjectives (plant-feeding, pupal, spine-like, self-made and worm-like), compared to the rare singleton nouns, were somewhat easier on the pretest and more difficult on the posttest.

There was no statistically significant main effect for writers or foil type nor was there a significant two-way interaction between writers and foil. This result is somewhat surprising in that different writers would be expected to write easier or harder items when they had the latitude to choose their own foils. Apparently, the method of writing foils, which allowed the choice of only single words to replace the question word, did not provide as much latitude as was expected. If the item writer were allowed to rewrite or replace the entire noun phrase, perhaps the differences between writers would have been magnified.

The lower half of Table 1 shows one interaction (RFS) involving foil type--the interaction of repeated measure, foil type and keyword vs. singleton. All of the posttest means in this portion of the table are very similar. A Newman-Kuels <u>a posteriori</u> test of the differences between pretest item difficulties in this interaction was calculated. This revealed that among the items with "writer's choice" foils, the rare singleton items were more difficult on the pretest (40.2%) than the keyword items (52.5%).

2. <u>Variance between writers</u>. Variability of item difficulties across item writers was examined to answer the question, "Do the

difficulties of items constructed with the 'writer's choice' foils vary more widely across writers than the difficulties of items constructed with the algorithmic foils?" The expectation was that some writers would choose very difficult foils for a given transformed sentence and other writers would choose very easy foils for that same sentence. In contrast, the algorithmic foils which were chosen at random from matched groups of similar words from the passage would seem to be free of any item-writer bias, and, hence, less variable in their effects on item difficulty.

In examining the variability across writers, the focus was on each sentence that was to be transformed by each writer. There were a total of 20 sentences taken from the passage on insect development. For example, each of the four item writers produced items from the sentence, "In most cases each egg produces a single immature insect," for which the keyword adjective was "immature." It was, therefore, possible to identify four item difficulties for a given combination of sentence and foil techniques. For instance, for the sentence containing the word "immature," each writer generated an item using the "writer's choice" foil method, and this resulted in pretest difficulties of 30%, 65%, 52%, and 37%. Posttest item difficulties for these items were 67%, 63%, 74%, and 52%, in that order. The pretest and posttest variabilities could then be calculated across these item difficulties.

After all variances of item difficulties across writers were calculated, they were subjected to a repeated-measures analysis of variance in which the dependent variables were the natural logarithms of the variances (Scheffé, 1959, p. 83). The design for this analysis was 2 x 2 x 2 x 2 with the following factors: (1) foil type (writer's choice

vs. algorithmic), (2) part of speech (noun vs. adjective), (3) stem type (keyword vs. rare singleton question word), and (4) the repeated measure (pretest vs. posttest).

The surprising result of this analysis was that there were no significant main effects or interactions at the .05 level of significance. For example, even though the writer's-choice foil method had an average variability of 115.31% (standard deviation of 10.74%) and the algorithmic foil method had a variability of 73.97% (s.d. of 8.60%), this difference was statistically significant at $p = .07$. The variabilities of item types are shown in Table 3.

Insert Table 3 about here

Apparently, the writer's choice method involved a good deal of constraint on the item writers who could only replace a single word with alternative words. Further research should examine the differences in variability when greater latitude is given to the item writers.

One important limitation of the present study that should be mentioned is that only four item writers were employed. Calculation of variabilities across only four writers is clearly susceptible to the influence of any one of the four item difficulties. With a larger sample of writers, the effects may have been more clearly detectable.

3. Errors of measurement. KR-20 internal consistency estimates of reliability and standard errors of measurement were computed. Four scales were examined: (a) noun-based, rare singletons (P1S1), (2) noun-based, keywords (P1S2), (3) adjective-based, rare singletons (P2S1), and

(4) adjective-based, keywords (P2S2). Median standard errors and the range of these standard errors are presented in Table 4.

Insert Table 4 about here

The practical significance of these differences among the first three is small indeed. The last combination of conditions, the adjective-based keywords, showed the largest degree of error over the eight test forms. This runs contrary to the criterion of instructional sensitivity examined earlier in this study, where the adjective-rare singleton (P2S1) and the noun-keyword (P1S2) conditions showed the lowest sensitivity to instruction.

The very same analysis occurred with two 10-item scales for each form representing the two foil conditions, writer's choice and algorithmic. The median standard errors and ranges for these are shown in Table 4. Thus the factor of foil method appeared to have little to do with the degree of measurement error present in these scales and test forms.

4. Concurrent validity. Six types of items were grouped and scored to create six variables: (a) a scale of rare singleton nouns (P1S1), (b) a scale of keyword nouns (P1S2), (c) a scale of rare singleton adjectives (P2S1), (d) a scale of keyword adjectives (P2S2), (e) a scale of items with writer's choice foils (F1), and (f) a scale of items with algorithmic foils (F2). Correlations were computed among these six scales, and with the criterion measures of Cloze scores and total test score (F1 + F2). These correlations are shown in Table 5.

Insert Table 5 about here

Tests of differences between correlations were conducted using Fisher's z-transformation, and several small differences were found. The scale composed of rare singleton adjectives (P2S1) had a significantly lower correlation (.20) with Cloze test score than any of the other scales. The keyword noun scale had the lowest correlation (.60) with total test score, but had a moderate correlation with Cloze test score (.32).

The 10-item scales for writer's choice (F1) and algorithmic foils (F2) correlated more highly with the criteria than the 5-item scales (P1S1, etc.) as expected. F2 correlated more highly with the Cloze test than F1, but F1 correlated more highly with total score than F2.

Thus, no striking differences seem to emerge in terms of the concurrent validity of various types of items, as judged by the small differences between the correlations in Table 5. Further replication of the findings on other samples would be necessary to establish the stability of these differences.

Conclusions

Item Writers

An important finding of this study is that no significant difference between item writers was found in the analysis of item difficulties. In contrast, the study by Roid and Haladyna (1978) found significant differences in item difficulties between item writers who wrote from learning objectives or loosely-defined rules for transforming sentences. Also, no significant differences in the variability of item difficulties across item writers was found for any of the versions of the algorithms used in

this study. Further research could compare the methodology of item writing used in this study with the objective-based and rule-based methods of Roid and Haladyna (1978). Such research could test whether the methods of the present study reduce the apparent item-difficulty bias created by differences between item writers who use learning objectives as a guide. In any case, the present study clearly shows the feasibility of the sentence-based method.

The concept of identifying high information words from prose passages for use in selecting sentences to be transformed into questions is apparently a workable methodology. Words which have a SFI of 60 or less (they are relatively rare in American textbooks) and yet occur only once or twice in a prose passage are apparently good candidates for questions that test learning from prose. In contrast, nouns which are rare in American textbooks yet highly frequent in a prose passage are apparently not very good candidates for question words as they are too simple even on a pretest. This is particularly true, apparently, if the first occurrence of the noun in the passage is used. Thus, the concept of "high information" words should be considered a feasible and useful basis for selecting important sentences from prose passages for transformation into questions which measure reading comprehension. However, keyword nouns, which are relatively rare in American textbooks but occur very frequently in a prose passage, should not be used as question words, particularly if the sentence in which they first occur is used to form the question.

Algorithmic methods of constructing foils for multiple-choice versions of sentence-derived questions as implemented in the present study were shown to be feasible. One disadvantage of these sentence-transformation methods may be that they are somewhat easy on a pretest.

However, they still appear to generate items that show significant instructional sensitivity--a shift in difficulty from pretest to posttest when instruction is given between testing sessions. Algorithmically-derived foils for items functioned approximately the same as foils produced by item writers. One limitation of this study may be that the "writer's choice" foil method allowed only a single word to be varied between item writers. Further studies could explore the effects of giving more latitude to the item writer.

The present study provides evidence to support the theory of processing in reading proposed by Finn (Note 1). Rare singleton nouns, as compared to keyword nouns, were shown to be much more difficult for subjects to identify as answer words. In other words, rare singletons are truly high information words. The present study further suggests that rare words that occur on two or three occasions in a passage may have the same attributes as rare singletons.

Suggestions for future research would include comparison of sentence-derived questions with (a) other discourse-derived items (Bormuth, 1970, p. 50), (b) objective-based items, (c) items generated from item forms (Hively, 1974; Millman, Note 4), (d) items derived from concept analyses (Anderson, 1972; Tiemann, Kroeker & Markle, Note 5), and other methods such as (e) items based on facet analysis (Engel & Martuza, Note 6). These various approaches to item writing, and those of the present study, represent a core of a technology for item writing. Further research, development, and application of this technology is clearly needed in a variety of subject matters and instructional settings.

Reference Notes

1. Finn, P. J. Word frequency, information theory and Cloze performance: A lexical-marker, transfer-feature theory of processing in reading. Unpublished paper, State University of New York at Buffalo, School of Education, 1977. (Submitted for publication in Reading Research Quarterly.)

2. Roid, G. H., Haladyna, T., & Finn, P. J. A comparison of methods for transforming sentences into test questions for instructional materials. Technical Report, Teaching Research, Monmouth, Oregon, 97361, February, 1978.

3. Haladyna, T. M., & Roid, G. H. The quality of domain-referenced test items. Paper presented at the meetings of the American Educational Research Association, San Francisco, April, 1976.

4. Millman, J. Creating domain-referenced tests by computer. Paper presented at the meetings of the American Educational Research Association, New York, April, 1977.

5. Tiemann, P., Kroeker, L. P., & Markle, S. M. Teaching verbally-mediated coordinate concepts in an on-going college course. Paper presented at the meetings of the American Educational Research Association, New York, April, 1977.

6. Engel, J. D., & Martuza, V. R. A systematic approach to the construction of domain-referenced multiple-choice test items. Paper presented at the meetings of the American Psychological Association, Washington, D.C., September, 1976.

References

Anderson, R. D. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Bloom, B. S. Learning for mastery. Evaluation Comment, UCLA, Vol. 1, No. 2, May, 1968.

Carroll, J. B., Davies, P., & Richman, B. Word frequency book. Boston: Houghton-Mifflin, 1971.

Cronbach, L. J. John R. Bormuth: On the theory of achievement test items. Psychometrika, 1970, 35, 509-511. (Book Review)

Culhane, J. W. CLOZE procedures and comprehension. The Reading Teacher, 1970, 23, 410-413.

Dale, E., & Chall, J. S. A formula for predicting readability. Educational Research Bulletin, 1948, 27, 11-28.

Finn, P. J. A question writing algorithm. Journal of Reading Behavior, 1975, 4, 341-367.

Fredericksen, C. H. Representing logical and semantic structure of knowledge acquired from discourse. Cognitive Psychology, 1975, 7, 371-458.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10.

Millman, J. Criterion-referenced measurement. In Popham, W. J. (Ed.) Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Company, 1974.

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1967.

Roid, G. H., & Haladyna, T. A comparison of objective-based and modified-Bormuth item writing techniques. Educational and Psychological Measurement, Spring, 1978, in press.

Scheffe, H. The analysis of variance. New York: Wiley, 1959.

Footnote

[1]Paper presented at the meetings of the American Educational Research Association, Toronto, March, 1978.

This research was supported by Contract No. MDA-903-77-C-0189, from the Department of Defense, Advanced Research Projects Agency. Conclusions and recommendations in this study are those of the authors and not necessarily those of DARPA. The authors wish to acknowledge the consultation of Jason Millman, John Bormuth and Patrick Finn in the design of this research. Also, the help of Joan Shaughnessy in conducting the experiments and in refining the algorithms is gratefully acknowledged.

Table 1

Cell and Marginal Means and Standard Deviations

of Statistically Significant Results

	Pretest (R1)		Posttest (R2)		Total	
	Mean	s.d.	Mean	s.d.	Mean	s.d.
Noun (P1)	50.8	20.1	80.3	14.1	65.6	14.7
Keyword (S1)	61.3	15.7	83.5	12.8	72.4	12.3
Rare Singleton (S2)	40.4	18.6	77.3	14.8	58.8	13.8
Adjective (P2)	44.3	18.2	68.4	21.5	56.3	18.1
Keyword (S1)	39.4	14.0	67.4	15.6	53.4	13.4
Rare Singleton (S2)	49.1	20.6	69.4	26.3	59.3	21.6
Keyword (S1)						
Writer's Choice (F1)	52.5	18.2	75.3	17.5	68.9	16.3
Algorithm (F2)	48.2	18.6	75.6	15.4	61.9	15.8
Rare Singleton (S2)						
Writer's Choice (F1)	40.2	20.2	74.0	21.4	57.1	17.9
Algorithm (F2)	49.3	19.0	72.7	21.9	61.0	18.1
Total	47.6	19.4	74.4	19.1	61.0	17.1

Table 2

Keywords in the Insect Development Passage

and Their Text Frequency

Keyword	Text	Keyword	Text
Nouns	Frequency	Adjectives	Frequency
insect	8	immature	3
Insects	20	incomplete	2
metamorphosis	9	nymphal	2
egg	8	aquatic	2
adult	8	distinctive	2

Table 3

Variabilities and Standard Deviations of Item Difficulties

Item Types		Pretest	Posttest	Total
Foil Type (F)				
Writer's Choice	var.	131.39	101.21	115.31
	s.d.	11.46	10.06	10.74
Algorithmic	var.	69.94	78.24	73.97
	s.d.	8.36	8.85	8.60
Part of Speech (P)				
Noun	var.	94.45	85.63	89.93
	s.d.	9.72	9.25	9.48
Adjective	var.	97.30	92.47	94.85
	s.d.	9.86	9.62	9.74
Item Type (S)				
Keyword	var.	102.19	87.06	94.32
	s.d.	10.11	9.33	9.71
Rare Singleton	var.	89.93	90.95	90.44
	s.d.	9.48	9.54	9.51

Table 4

Median Standard Errors and Ranges for Test Scales

Composed of Various Item Types

	Median	Range
Rare Singleton Nouns (P1S1)	.915	.81-.96
Keyword Nouns (P1S2)	.870	.75-.99
Rare Singleton Adjectives (P2S1)	.930	.80-.99
Keyword Adjectives (P2S2)	.985	.94-1.07
Writer's Choice Foils (F1)	1.300	1.13-1.42
Algorithmic Foils (F2)	1.320	1.26-1.45

Table 5

Means, Standard Deviations and Correlations Between Scales Based on Item Types and Criterion Measures

	Mean	s.d.	Cloze	P1S1	P1S2	P2S1	P2S2	F1	F2	Total
1. Cloze Test Score (C)	16.9	5.4								
2. Rare Singleton Noun (P1S1)	4.1	1.3	.32							
3. Keyword Noun (P1S2)	4.2	1.0	.32	.32						
4. Rare Singleton Adjective (P2S1)	3.5	1.1	.20	.32	.16					
5. Keyword Adjective (P2S2)	3.3	1.3	.28	.27	.16	.16				
6. Writer's Choice Foils (F1)	7.4	1.9	.33	.49	.39	.57	.66			
7. Algorithmic Foils (F2)	7.6	1.7	.38	.57	.59	.44	.35	.28		
8. Total Test Score (Tot)	15.0	2.9	.44	.66	.60	.63	.65	.83	.77	

Note: Correlations are based on N=151 subjects who had complete Cloze test data. The Cloze test had 50 items, P1S1-P2S2 had 5 items each, F1 and F2 had 10 items and Total Test Score had 20 items.